How to design and analyse cluster randomized trials with a small number of clusters?

International Journal of Epidemiology, 2018, 998–1001
doi: 10.1093/ije/dyy061
Advance Access Publication Date: 18 April 2018
Letters to the Editor

Letters to the Editor

How to design and analyse cluster randomized trials with a small number of clusters? Comment on Leyrat *et al.*

Gerard JP van Breukelen[1,2]* and Math JJM Candel[1]

[1]Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands and [2]Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

*Corresponding author. Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: gerard.vbreukelen@maastrichtuniversity.nl

Cluster randomized trials (CRTs) are popular in public health and primary care, among other arenas, and for good reasons. If individual randomization is impossible or leads to serious treatment contamination, a CRT is the next best option as it shares with individually randomized controlled trials (RCTs) the prevention of confounding.

Unfortunately, CRTs have lower power than RCTs due to the design effect (DE), that is the inflation of the sampling variance (i.e. the squared standard error) of the treatment effect by intraclass correlation (ICC). Further, the data from a CRT must be properly analysed, taking into account the DE plus cluster size variation and the correct degrees of freedom (*df*) for testing the treatment effect. This in turn has consequences for the sample size calculation for a CRT, which must also take into account these three factors.

Using a simulation study, Leyrat *et al.*[1] compare the type I error rate and power of various analysis methods for CRTs with a quantitative outcome under various conditions concerning the ICC, the number of clusters, the average cluster size and the amount of cluster size variation. From their results they derive practical advice about the best methods of analysis, and they suggest using simulations to adjust the sample size for the lower-than-nominal power of these best methods in the case of small number of clusters.

The results of their study are important and we fully support their aims. However, we would like to point out that there is a simple alternative to their recommendation to use simulations to adjust the sample size for the lower-than-nominal power, and a better method than theirs to adjust for cluster size variation. Based on publications that appear to have been overlooked by Leyrat *et al.*, we first explain why their best methods of analysis are best indeed, and then show how to adjust for cluster size variation, and how the lower-than-nominal power for small number of clusters can be solved without simulations. We first summarize their results and then elaborate our comment.

Leyrat *et al.* evaluate four methods of analysis based on cluster means (weighting by cluster size, weighting by inverse variance of the cluster mean, parametric unweighted analysis, non-parametric unweighted analysis), and eight methods based on individual data, but taking the clustering into account [mixed regression with five methods to determine the df for the treatment effect test, and generalized estimating equations (GEE) with model-based standard error (SE), robust SE, or robust SE with small sample correction]. In all conditions, the nominal type I error rate and nominal power are 5% and 80%, respectively. They find that, especially for small number of clusters, the type I error rate is seriously inflated by cluster mean analysis weighted by cluster size, mixed regression without correction for small df, and GEE without small sample correction, whereas all other methods then suffer from lower-than-nominal power.

Close inspection of the figures in their online supplement shows that cluster means analysis weighted by inverse variance, and two versions of mixed regression with corrected df, perform best, with an actual power between 70% and 80% for a total of 20 clusters and between 60% and 80% for 10 clusters, depending on the other factors in

their simulation study: ICC, average cluster size, and coefficient of cluster size variation (CV). GEE with small sample size correction performed similarly, except for very small ICC and large CV. All other methods with proper type I error rate had a lower power than these methods. These results are reflected in Leyrat *et al.*'s summary Table 2 of recommended methods of analysis. Further, the authors recommend using simulations to adjust the sample size for the lower-than-nominal power of these best methods (see their Discussion).

We believe that many results in Leyrat *et al.* can be understood by looking at a few publications in statistical journals, and that there is a quicker and easier way to manage sample size calculation for CRTs than simulations. The following statistical results are relevant: (i) the relation between analysis of cluster means and mixed regression; (ii) the optimality of weighting cluster means by inverse variance; (iii) the effect of cluster size variation on the required sample size; and (iv) correcting the sample size calculation for small df.

First, we comment on the relation between analysis of cluster means and mixed regression. As shown in our 2003 paper,[2] if all clusters have the same size in the sample (so CV = 0), then unweighted analysis of cluster means is equivalent to mixed regression of the individual data taking the clustering into account. If clusters vary in size, then weighting cluster means by inverse variance is equivalent to mixed regression, at least for large samples.[3] For small samples, the two methods may differ somewhat, depending on how variance components are estimated.

Now we deal with the issue of weighting cluster means if clusters do vary in size. As shown previously,[3–5] weighting cluster means by inverse variance is more powerful than unweighted analysis (especially for an ICC near zero), and also more powerful than weighting by cluster size (especially for an ICC larger than the inverse mean cluster size). The lower power of unweighted analysis of cluster means is visible in most figures of Leyrat *et al.*'s supplement. The lower power of weighting by cluster size is not visible in their figures, but this method has an inflated type I error in those figures which correspond to large CV and ICC larger than the inverse mean cluster size (see their online figures 8, 11, 12). This may be due to using an incorrect standard error. Adjusting for the inflated type I error risk, whether by SE correction or lowering α, will inevitably lower the power of cluster size weighting.

Next, we clarify the issue of accounting for cluster size variation in the sample size calculation. As shown elsewhere,[4–6] the power loss due to cluster size variation can be restored by increasing the number of clusters with a percentage that depends on the CV of cluster size as well as on mean cluster size and ICC through a simple mathematical equation. As we have shown,[5] this percentage never exceeds 100%*(CV^2/2) or 100%*[CV^2/(4-CV^2)], depending on which of two mathematical approximations we use, the first always giving an overadjustment and the second sometimes a slight underadjustment. For the CVs in Leyrat *et al.*, this gives about 8% or 4% extra clusters if CV = 0.4 (small), and 32% or 19% extra clusters if CV = 0.8 (large), depending on which of the two approximations is used. In their appendix (page 10, last equation), Leyrat *et al.* also adjust their sample size for cluster size variation, apparently using an approximation from Eldridge *et al.*[7] However, as correctly stated in Eldridge *et al.*,[7] that adjustment is based on analysis of cluster means weighted by cluster size. That method is inefficient if the ICC is larger than, say, the inverse mean cluster size and the CV is large,[5] and it can lead to almost 100%*CV^2 extra clusters, which is twice as much as the overadjustment based on our work.[5] So Leyrat *et al.* overadjust their sample size especially for large ICC and CV, which correspond to their supplementary figures 8, 11 and 12 where cluster size weighting has an inflated type I error rate instead of the expected lower power (remember that adjusting that analysis method to get the correct type I error rate will lower its power). Incidentally, there is a typo in the equation in Leyrat *et al.* (page 1293) (not in Eldridge *et al.*[7]), causing the DE to be (1-ICC) if CV = 0, whereas the correct DE is 1+(m-1)*ICC if CV = 0, where m = cluster size (see elsewhere[7–9]). Further, the adjustment for unequal cluster sizes can be applied before instead of after rounding the number of clusters as computed with classical sample size equations such as in our paper[8] upward to the nearest integer. For instance, if the classical computation gives 8.3 clusters per arm and we need to increase that by 8%, then we may first multiply 8.3 with 1.08 to get 8.964 clusters which is then rounded to 9 clusters per arm. If we first round and then increase with 8%, we get 9*1.08 = 9.72 clusters, rounded to 10 clusters per arm.

Finally, there is the df needed for sample size calculation. As shown elsewhere,[10,11] the power loss due to using the t-distribution with the correct df in data analysis, if the sample size has been calculated with the standard normal distribution, can be compensated by adding two clusters per treatment arm. This holds for a nominal power of 80% as well as of 90%, provided the type I error risk is set at 5% and the number of clusters per arm according to the standard normal approximation is at least eight (for less than eight clusters per arm add three clusters per arm; for a 1% risk always add four clusters per arm). These results agree with those in Leyrat *et al.* According to their supplement, the actual power of mixed regression with df = k-2 (their between-within correction) varied from 60% to 70% for a total of k = 10 clusters (i.e. five per arm). From this actual

power we can compute the number of extra clusters needed to have an actual power of 80% in a simple way, as follows. In all sample size equations for two-arm trials, whether RCT[12,13] or CRT,[8] the sample size is proportional to a term $(\text{tpower} + \text{talpha})^2$. Here, tpower is the (1-beta)-th percentile of the Student t-distribution for a power (1-beta), and talpha is the alpha/2-th percentile of that distribution for a type I error risk alpha if we test two-tailed. For instance, if $k = 10$ so that $df = 8$, then tpower = 0.89 for 80% power and talpha = 2.31, giving $(\text{tpower} + \text{talpha})^2 = 10.24$. If the actual power is 60% instead of 80%, then tpower = 0.26, giving $(\text{tpower} + \text{talpha})^2 = 6.60$. The ratio 10.24/6.60 is 1.55, which means that we need to multiply k with a factor 1.55 to get an actual power of 80%. Given $k = 10$ clusters, we thus need to increase k to 16, which is three extra clusters per treatment arm. Similar calculations for other k in Leyrat's simulations, taking the actual power of mixed regression with $df = k\text{-}2$ from their figures, also lead to two or three extra clusters per arm, as recommended elsewhere.[10,11] In short:

i. the superior performance of weighting cluster means by inverse variance and of mixed regression with proper df follows from results in statistical literature;
ii. the power loss arising from cluster size variation can be compensated by adding clusters following simple approximations in our paper[5]; and
iii. we do not need simulations to find out how many extra clusters we need in a CRT with a small number of clusters to compensate the power loss arising from the difference between a z-test and t-test with small df; we simply add two or three clusters per treatment arm if alpha = 5% two-tailed, or four if alpha is 1% two-tailed.

To this summary we add a few notes in response to questions by the reviewer of this letter. First of all, researchers are advised to plan at least 10 clusters per treatment arm for two reasons. One reason is the fact that non-normality of the cluster effect can invalidate the significance testing and confidence interval for the treatment effect, especially if the number of clusters is small. As the number of clusters goes up, the central limit theorem in statistics ensures approximate normality of the treatment effect estimate even if the cluster effect is not normally distributed. The other reason is that the power of a cluster randomized trial with fewer than 10 clusters per arm will often be too low. For instance, for a medium effect size $d = 0.50$, where d is Cohen's d,[14] a two-tailed alpha of 5% and a power of 90%, we need 86 persons per treatment in a classical RCT. In a cluster randomized trial with a typical ICC of 0.05 and a sample size of 20 persons per cluster, the design effect (DE) is 1.95, implying a sample size of $1.95*86 = 168$ persons per arm, giving 8.4 clusters per arm. Even ignoring cluster size variation, but taking into account the df adjustment discussed in this letter, we thus need at least 11 clusters per arm. One might lower this by accepting a power of 80% (and so a type II error risk of 20%!), but cluster size variation and effect sizes smaller than 0.50 are omnipresent in health research, and both call for an increase of the number of clusters.

A second note concerns cluster randomized trials with a binary instead of quantitative outcome. For binary outcomes, sample size calculation with an adjustment for varying cluster size is explained and demonstrated in our work elsewhere,[15] based on mixed logistic regression. However, the issue of the correct df has not been explored yet. The analysis of binary outcomes is usually based on Wald or likelihood ratio tests, both involving the standard normal instead of t-distribution, and assuming fairly large samples.

As a last note, the issues in this letter also arise for other nested designs, of which we here mention two. For multicentre trials (with centre as random effect), equations for sample size calculation and adjustments for varying sample size per centre are presented elsewhere.[4,16–18] For stepped wedge cluster randomized trials, things are more complex because of the confounding between treatment and period that has to be adjusted for, and because allowing for treatment by period interaction can easily lead to unidentifiable models. There are useful references for sample size planning of stepped wedge cluster randomized trials assuming a constant treatment effect.[19–22]

Conflict of interest: None declared.

References

1. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol* 2018;**47**:321–31.
2. Moerbeek M, Van Breukelen GJP, Berger MPF. A comparison between traditional methods and multilevel regression for the analysis of multi-center intervention studies. *J Clin Epidemiol* 2003;**56**:341–50.
3. Searle S, Pukelsheim F. Effect of intraclass correlation on weighted averages. *Am Stat* 1986;**40**:103–05.
4. Van Breukelen GJP, Candel MJJM, Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med* 2007;**26**:2589–603.
5. Van Breukelen GJP, Candel MJJM. Efficiency loss due to varying cluster size in cluster randomized trials is smaller than literature suggests. *Stat Med* 2012;**31**:397–400.
6. Candel MJJM, Van Breukelen GJP, Kotova L, Berger MPF. Optimality of unequal cluster sizes in multilevel studies with

realistic sample sizes. *Commun Stat Simul Comput* 2008;**37**: 222–39.
7. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006;**35**:1292–300.
8. Van Breukelen GJP, Candel MJJM. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! *J Clin Epidemiol* 2012;**65**:1212–18.
9. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004;**1**:80–90.
10. Lemme F, van Breukelen GJP, Candel MJJM, Berger MPF. The effect of heterogeneous variance on efficiency and power of cluster randomized trials with a balanced 2x2 factorial design. *Stat Methods Med Res* 2015;**24**:574–93.
11. Candel MJJM, Van Breukelen GJP. Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Stat Methods Med Res* 2015;**24**:557–73.
12. Julious SA. *Sample Sizes for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC, 2010.
13. Kirkwood BR. *Essentials of Medical Statistics*. Oxford, UK: Blackwell, 1988.
14. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Mahwah, NJ: Erlbaum, 1988.
15. Candel MJJM, Van Breukelen GJP. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat Med* 2010;**29**:1488–501.
16. Moerbeek M, Van Breukelen GJP, Berger MPF. Design issues for experiments in multilevel populations. *J Educ Behav Stat* 2000;**25**:271–84.
17. Moerbeek M, Van Breukelen GJP, Berger MPF. Optimal experimental design for multilevel logistic models. *J R Stat Soc Ser D Stat* 2001;**50**:17–30.
18. Raudenbush SW, Liu X. Statistical power and optimal design for multisite trials. *Psychol Methods* 2000;**5**:199–213.
19. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28**:182–91.
20. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed models. *Stat Med* 2016;**35**:2149–66.
21. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomized trials. *Stat Med* 2016;**35**:4718–28.
22. Thompson JA, Fielding KL, Davey C, Aiken AA, Hargreaves JR, Hayes RJ. Bias and inference from misspecified mixed-models in stepped-wedge trial analysis. *Stat Med* 2017;**36**:3670–82.

Response to: How to design and analyse cluster randomized trials with a small number of clusters? Comment on Leyrat *et al.*

International Journal of Epidemiology, 2018, 1001–1002
doi: 10.1093/ije/dyy062
Advance Access Publication Date: 18 April 2018

Clémence Leyrat,[1]* Katy E Morgan,[1] Baptiste Leurent[1] and Brennan C Kahan[2]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK and [2]Pragmatic Clinical Trials Unit, Queen Mary University of London, London, UK

*Corresponding author. Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT London, UK. E-mail: clemence.leyrat@lshtm.ac.uk

We would like to thank Van Breukelen and Candel for their comments on our manuscript.[1] Although we broadly agree with them, we would like to clarify several points.

First, they argue that our results can be understood in light of the existing literature. We agree that some of the results in our article are well known (e.g. that unweighted cluster-level analyses lose efficiency). However, these approaches are still commonly used,[2] and so we included them in order to empirically demonstrate the benefit of other approaches. Furthermore, we are unaware of any empirical comparison between generalized estimating equations (GEEs), mixed-effect models and cluster-level analyses for continuous outcomes. We agree with Van Breukelen and Candel that some theoretical results are available for these approaches; however, these are often based on approximations which do not always translate to realistic scenarios (particularly regarding small-sample corrections), and so it is useful to assess the properties of these approaches across a range of realistic scenarios using simulation.[3]

Second, Van Breukelen and Candel take issue with the sample size formula used in our simulation study. Because sample size formulas depend on the underlying analysis model, there is no single formula which is appropriate for all the analysis methods being compared. However, our aim was to benchmark the relative performance of each analysis method in terms of type-I error rate and power. Given that the specific sample size formula used will have no impact on which analysis approach performs best, we are unsure why